关系数据库中实体解析研究综述*

高广尚[1,2] 张智雄[1]
（1. 中国科学院文献情报中心，北京 100190；2. 中国科学院大学，北京 100190）

摘要：[目的]分析关系数据库中实体解析技术的研究现状和未来研究方向。**[方法]**从实体解析的精度和效率两方面展开系统研究。精度方面基于增量式、统计方法和相关信息。效率方面基于分块、字符串相似和其他思想。**[结果]**最大化实体解析精度和解析效率是实体解析技术研究的主要目标，但在数据源的动态演化、异构性和非精确字符串匹配等方面的研究仍面临重大挑战。**[局限]**仅从实体解析过程中所需的精度和效率方面进行探讨，对解析模型本身的特点和局限性关注不足。**[结论]**本研究有助于更全面地了解关系数据库中实体解析的过程、研究现状和未来研究方向。

关键词：实体解析；记录链接；关系数据库
中图分类号：TP393

Survey on Entity Resolution over Relational Databases

Gao Guang-shang[1, 2] Zhang Zhi-xiong[1]
（1. National Science Library of CAS, Beijing 100190
2. University of Chinese Academy of Sciences, Beijing 100190）

Abstract: [Objective] To analyse the current status and future research direction of Entity Resolution(ER) over relational databases. **[Methods]** Systematical researches are conducted on two aspects of the accuracy and efficiency of ER. The accuracy of ER was based on incremental methods, statistical methods and related information. The efficiency of ER was based on blocking, string similarity and others. **[Results]** Maximizing precision and efficiency is the main objective of entity resolution, but research on dynamic evolution and heterogeneous of data sources and inexact string matching still faces significant challenges. **[Limitations]** Only precision and efficiency needed in the process of entity resolution are discussed, but the characteristics and limitations of ER model need to pay more attentions. **[Conclusions]** This paper will facilitate to give a comprehensive overview of the process of ER over relational databases, research status and future research direction.

Keywords: Entity resolution; record linkage; Relation Databases

1 引言

实体解析（Entity Resolution，ER）的研究已有很长一段时间，一些早期的研究工作可以追溯到20世纪50年代[1]，但现在它依然是一个活跃的研究领域。早在1969年，Fellegi和Sunter就基于指向同一现实世界实体的不同记录应具有某些共性这一假设，提出了一种链接记录的技术[2]，或称为实体解析技术，数据库领域的后续研究也大都遵循这一假设。实体解析技术已在各种名称下被广泛研究，包括记录链接（Record Linkage）[3]，合并/清洗（Merge/Purge）[4]，重复数据删除（Deduplication）[5]，参考协调（Reference Reconciliation）[6]，对象识

[1]*本文系国家“十二五”科技支撑计划课题“科技知识组织体系共享平台建设”（项目编号：2011BAH10B03）

别（Object Identification）[7]和其他等[8-10]。在关系数据库中，实体解析技术主要是指解析出描述同一实体的 n 个（n>1）相似重复记录，这里的记录又被称为数据。其中，解析模型可大致分成以下 3 类：基于匹配的聚类（基于布尔规则匹配信息的记录聚类）[4, 11]，基于距离的聚类（基于相对距离的记录聚类）[12, 13]或成对实体解析（Pairs ER，逐对解析记录）[5, 10, 14]。

随着“大数据”时代的到来，ER 技术在数据清理、数据集成和数据挖掘等研究领域中起着关键的作用，因此，在数据质量和信息共享等方面 ER 技术被视为一种重要的保障性技术。由于 ER 技术的应用范围涵盖了多个领域，例如，人口普查[15]、公共卫生、Web 搜索、商品列表比较、反恐、垃圾邮件检测和机器阅读等领域，因此，其引起了来自学术界、工业界很多专家的关注。尽管在结构化数据库中实施抽取、匹配和解析的 ER 技术是一项较为成熟的技术，但其面临的最具挑战性问题仍然是高效性和准确性，尤其是在复杂的大数据情况下。

国内外已有研究对实体解析过程的每个步骤所涉及的方法进行了介绍和阐述[9, 16]，但并未从实体解析目标这一角度来探索其解析策略。鉴于此，笔者从实体解析过程中所涉及的精度和效率两方面，来分析、整理相关文献以对关系数据库中实体解析研究进行综述。同时，期望该综述为将来实体解析研究的优化整合与进一步挖掘提供一些有价值的借鉴和参考。

2 实体解析技术概述

为对关系数据库中 ER 技术有一个简明扼要的完整性了解，本节将从实体与记录间的关系、实体解析过程两方面对其进行概述，使之形成一个相对完整的理论体系。

2.1实体与记录间的关系

所描述的实体可能是一个物理对象（例如一个人或一座房屋），或者可能是一个逻辑结构（例如一个家庭、一个社交网络或喜欢某一特定类型音乐的人物列表），他们被视为属于某个类别的集合，比如人物类别下的多个个体组成的集合。关系数据表是这些集合的数字表示。关系数据表包含一系列的记录（records）或条目（entries），其中，每一个记录与现实世界中的一个或多个实体相关联。特别地，每个记录可能指向一个特定的实体，但是每个实体可能有一个或多个描述它的记录，如图 1 所示。其中，记录是由列（属性、域）组成。很显然，所有记录的模式结构是相同的，这有利于实体解析算法的应用。

Tom
Ann
Bill

NAME	STREET	BIRTHDAY	SALARY
Tom	17 Lexington Ave.	24th Jan, 1948	$80,000.00
Tom	55 Lexington Ave.	24th Jan, 1948	$80,000.00
Ann	5 5th Ave.	9th Feb, 1982	$70,000.00
Bill	3 3rd Ave.	26th Oct, 1974	$60,000.00

现实世界中实体　　数字表示

图 1 现实世界中实体和相应的数字表示

特别地，图 1 中有 3 个实体（人物）被 4 个记录所表示，其中有 2 个重复的记录（NAME 属性值为“Tom”的两个记录），本文将他们称为相似重复记录。

通常情况下，人们希望删除关系数据表中的相似重复记录，因为这些相似重复记录是影响数据质量的关键问题之一，比如在数据集成等系统中。为此，可采用的较合理处理方式有：或者将相似重复记录合并成单个记录，或者链接每个相似重复记录。最后，正如 M LLER H 等[17]在研究中指出，解析这些重复的记录以让他们指向同一实体的任务不仅非常艰巨，而且可能非常有意义。

2.2实体解析过程

在关系数据库中，实体解析过程通常分为三个阶段：预链接阶段、链接阶段和后链接阶段[9, 18]。

（1）预链接阶段。记录被预处理或规范化，以提高链接精度。预链接阶段可能是这 3 个阶段中最依赖于上下文的阶段，因为其目标是转换记录的属性数据以使链接操作尽可能地容易实现。潜在的转换操作可能是将日期、电话号码和地址等属性中的数据转换成标准格式表示的数据，或在数据集合中拆分/合并以匹配另一个模式。

（2）链接阶段。执行实际的记录链接。链接阶段包括确定性链接和概率性链接[19]。

确定性链接涉及一个或多个属性的精确匹配，它又分为①简单确定性链接和②传递确定性链接。简单确定性链接的思想是：如果多个记录在给定对应属性中具有相同的值，就将这些记录链接起来。简单确定性链接是最简单的方法。传递确定性链接的思想是：如果多个记录存在有任何属性值匹配，就将这些记录链接起来。传递确定性链接能在某些属性值缺失情况下推理出多个记录指向同一实体。

概率链接又称为模糊链接，即在两个属性值不完全匹配的情况下，对记录实施链接。其中，最著名的是 Fellegi-Sunter 模型[20]，在属性条件独立的情况下，该模型描述了一组被证明是最优的规则。其主要思想是计算每个属性的区别能力（Discriminatory Power），然后组合这些属性以得到一个判定两个记录是指向同一实体的概率。然而，在大数据集中，概率链接不易实现。此外，由于概率链接方法考虑匹配项（记录或属性）的概率，因此，在减少 false non-matches 的同时（被错误的分为不匹配的记录），也可能会产生 false matches（被错误的分为匹配的记录）。

（3）后链接阶段。审查链接的结果，检查“可能链接”的记录，并最终使用这些结果。该阶段存在 3 种操作方式：①链接，②合并和③相似重复记录删除链接：在每个记录中增加一个指向其他链接记录的引用，或在不同的数据集中存储这些链接。合并：如果要链接两个不同的数据集，可以在它们之上创建一个使两个数据集统一的视图。相似重复记录删除：仅保存一条必要的记录以避免重复。

3　以精度为目标的实体解析

解析精度主要涉及记录间比较技术的质量，这主要体现在解析过程中所采用的记录比较方式。根据方法的不同，当前的研究主要有基于增量式、基于统计方法和基于相关信息 3 类方法。

3.1基于增量式

与在解析过程中假设规则和数据固定不变这一情形不同的是，基于增量式

方法的实体解析过程将规则和数据将视为动态变化的，因此能很好地适应具有复杂结构、数据更新速度快的大数据环境。

（1）规则演化

SE Whang 等[21]针对实体解析结果相互影响的问题，基于解析规则的动态语义和解析规则之间的关系，提出了一种针对解析规则变化情况的实体解析方法。该方法考虑如何利用已有的解析结果来深入地研究实体解析问题。特别地，对规则演化进行了形式化，提出了规则单调和上下文无关两个约束，指出满足这两个约束的规则可使用增量方式进行处理。由于采用新规则时的解析过程能利用先前的解析结果，因此能减少计算复杂度，并提高解析精度。

Steven Euijong Whang 等[22]指出实体解析过程不是一次性过程，而是一个随着人们对数据、模式和应用的认识程度的加深而变化的过程。在大多数情况下，用来解析记录的逻辑规则会不断演变，因为应用本身会不断演变，而且用于比较记录的专业知识水平也会不断提高。由于将这些变化因素考虑进去，因此解析精度能不断得到提高。特别地，作者认为在对大规模数据集进行解析时，所采用的从头开始重新进行解析的朴素方法（naive）是不能容忍的，因为计算代价高昂。

Steven Euijong Whang 等[23] 针对演化规则提出了一个增量式实体解析方案。由于该方案借助迭代块和联合实体解析两种方法，因此它能提供很好的扩展性和精确性，并能适用于不同的应用领域。

（2）数据演化

通常，实际中采用的数据分块方法并不能保证块间数据的独立性，因为有些相似的记录可能被分配到不同的块中。在这种情况下，分块方法在提升解析效率的同时，也降低了解析精度。为解决这个问题，Steven Euijong Whang 等[24]基于增量计算的思想，提出了迭代的实体解析方法。在每次迭代中，首先把上一次迭代计算得到的每个分块的实体解析结果传输到其他块内，然后每个分块根据收到的更新结果增量式地计算各自块内的实体解析结果，这样的迭代计算一直进行直到结果不再改变或迭代次数达到给定阈值。该方法在保证解析效率的前提下提高了解析结果的精度。

Anja Gruenheid 等[25]注意到大数据时代下的数据更新速度往往较快，这将使得以前的解析结果很快失效。为解决此问题，作者提出了一个端到端的框架，它能在数据更新（包括插入，删除和修改）到来时以一种增量式方法更新解析结果。重要的是，在不影响原有解析结果的情况下，提出的算法不仅能将数据更新中的记录与现有的聚簇进行合并/分离，还能利用数据更新中的新证据来修正先前存在的解析错误。实验表明，算法能显著地减少解析时间，同时无损解析质量

Sunita Sarawagi 等[26]从另外的角度出发，针对 top-k 计数查询提出了“一边求解查询一边解析实体”的方法。算法的基础在于，一般的查询涉及的数据记录数量较小，算法没有必要在所有数据记录上运行实体解析算法，仅需要处理查询结果中涉及到的记录。该方法的难点在于，解析查询结果中的记录可能需要查询结果之外的数据记录，而快速得到查询结果以外的相关数据记录也是一件困难的事情。

Benjelloun 等[11]提出了“F-Swoosh”算法，该算法能很好地适应数据增量式的情况，且考虑到了新增的数据或特征。Heiko Müller 等[17]认为清洗数据是一项耗时且代价高昂的任务。在已获得干净的数据集合后，当数据集中的一个记录值出现更改时，清洗过程仅需从包含该更改值的记录开始即可，从而避免对整个

数据库执行清洗的过程。Hernandez 等[27]认为在对数据进行合并和清洗前，串联所有数据所需的时间和空间被证明是代价高昂的。为此，提出了一个增量式算法在短时间内能很好地解析新增的数据。

另一个与增量式解析技术密切相关的是增量式图形聚类。Claire mathieu 等[28]研究增量式相关性聚类（Incremental Correlation Clustering）。作者认为，当数据源不断动态变化、演化时，每次数据更新操作都从最开始应用解析的方法是代价高昂的。为解决速度方面的问题，作者采用增量式聚类技术。其中，算法主要关注两点：①每次增加一个结点；②已识别的聚类结果需要保存。Charikar, M.等[29]研究增量式聚类，与其他聚类方法不同的是，该方法需要预先设定聚类结果中聚簇的数目。该聚类方法的思想是：在给定数据流的情况下，增量式聚类算法使形成的最大聚簇直径最小。最后，作者将增量式聚类问题定义为：对一个包含n个结点（数据）的更新序列来说，维持一个包含k个聚簇的集合，使得每当有输入结点出现时，它或者被分配到当前集合中的某一个聚簇中，或者在该集合中新增一个仅包含该结点的单元素聚簇。

此外，由于针对数据演化的实体解析问题与聚类数据流的问题密切相关，Aggarwal 等[30]提出了一个 CluStream 算法，由于考虑到数据流具有不断随时间变化的特性，因此提出的算法能在不断变化环境中的不同时间区间上很好地进行聚类操作。

3.2基于统计方法

与统计方法相关的是特征选择问题，特征选择的好与坏，直接决定了解析的精度。尽管统计方法增加了推理和学习的复杂度，但是通过利用这些以前被忽视的数据属性，可有效地改进传统的实体解析算法，从而提高解析精度。

Xin Dong 等[6]研究利用记录间的 3 个主要特征来实现一种有效利用机器学习的实体解析算法。首先，利用记录间的关联来为记录间的比较设计新方法。接着，传播记录的决策信息（匹配或不匹配）以累积正面和负面证据。最后，通过合并各属性值来逐步丰富各记录信息，从而提高了实体解析精度。Parag Singla 等[31]提出了联合推理方法，对所有候选匹配对进行同时推理，并允许信息从一个候选匹配对经由它们共有的属性传播到另一候选匹配对。由于该方法基于条件随机场（Conditional Random Fields，CRF），因而提高了实体解析精度。

此外，在基于统计学的实体解析方法中，参数设置错误和训练数据缺失会导致检测结果不准确。针对这类问题，Peter Christen 等[32]提出了一种两阶段的统计学方法。在第一阶段，从参与比较的记录对中自动选择高质量的训练样例，在第二个阶段，使用这些训练样例来训练一个支持向量机（SVM）的分类器。由于这种两阶段方法能有效地调整解析过程，从而能提高实体解析精度。

楼俊杰等[33]在基于马尔科夫逻辑网络（Markov Logic Networks，MLNs）的实体解析算法体系中，引入一个可变权重的规则，试图解决原有系统无法处理的记录二义性问题（两条记录中出现的“John Smith”其实并非指向同一人）。由于引入了更能反映实现情况的可变权重规则，因此提出的算法能在一定程度上提高解析精度。

3.3基于相关信息

尽管传统上实体解析算法通常使用各种属性相似措施来单独地匹配记录，但如果能利用其他相关信息来辅助实体解析过程，将使实体解析算法能很好地

适应大数据集环境，并使其具有很好的扩展性、灵活性。

Surajit Chaudhuri 等[34]通过挖掘文档集合，并利用参考实体表中每个实体的多个变体形式来扩展给定的参考实体表，这样一来就构成了一个字符串等价关系词典。由于能利用词典的精确信息来计算实体之间的相似性，因而提高了实体解析的精度。Liangcai Shu 等[35]提出了一个能描述实体之间关系的生成式潜在主题模型，隐含狄利克雷分布双主题模型（LDA-dual model），并给出了高精度的实体解析算法。由于该模型能使用语料库中全局信息来学习一个高性能的分类器，因而能提高解析精度。

Vibhor Rastogi 等[36]提出了一个扩展的实体解析算法，它能利用比较的中间结果来进行综合推理。由于该算法不仅能利用记录间的相似信息、记录同现的频度信息，还充分考虑了记录比较结果之间的影响，因而能提高实体解析精度。

3.4研究方法分析比较

以精度为目标的实体解析过程，主要关注相似重复记录间比较技术的质量。主要研究方法的比较情况，如表 1 所示：

表 1 以精度为目标的主要研究方法的比较

以精度为目标的实体解析（主要关注相似重复记录间比较技术的质量）		
采用的方法	**优点**	**缺点**
基于增量式：①规则演化；②数据演化	①有效地重复利用先前的解析结果不仅能提升效率，而且能提高解析精度；②采用的聚类算法是一种无监督学习算法，能够辅助相似性计算，从而能很好地应用于相似重复记录的解析；③利用关联结果迭代查找相似重复记录；④可以获得非常高的准确性。	①缺乏一定的灵活性；②解析优化过程所需时间开销大
基于统计方法：①条件随机场；②基于统计学和③马尔科夫逻辑网络等	①易于扩展；②对小数据集的处理效率较高，但随着数据规模的扩大，效率往往不能进一步提升，	①人工标注代价过高；②计算复杂度高；③参数难以确定，容易出现过拟合现象；④参数的确定依赖领域知识；⑤缺少特定的标准
基于相关信息：①等价关系词典；②语料库中全局信息和③相似信息、同现信息等	①不依赖于特定应用领域；②考虑隐藏在词汇背后的属性间的关联关系；③能搜索对应的字段匹配函数，从而避免采用固定的匹配函数应用于不同数据源可能引起的匹配精度波动较大的问题。	①引入不确定性；②使用的结构复杂；③一般夹杂着噪音，导致解析结果准确率较低

4　以效率为目标的实体解析

解析效率主要涉及解析算法的执行速度，这主要体现在两个方面：一是减少需要的记录对比较次数；二是提高记录属性值的比较效率。根据方法的不同，当前的研究主要有基于分块、基于字符串相似和基于其他思想等方法。尽管很多研究学者在提升解析效率方面做出了巨大努力，但现有算法在最坏情况下的时间复杂度仍为$O(n^2)$[37]，即计算复杂性仍然远超过线性，因而难以应用于大数

据。

4.1基于分块

当需要比较的记录规模较大时，传统上采用的基本技术是利用“嵌套”循环方式来逐一比较记录对，这将需要大量的计算开销。分块方法的目的就是为了缩小比较空间，进而减少记录间的比较次数，最终实现不影响解析准确性和完整性的较高解析效率目标。笔者从属性值、自动学习和分块方法比较3个方面对现有研究进行综述。

（1）属性值

为提升实体解析的效率，Hernández MA等[27]较早地提出了数据分块处理的思想。首先，记录按照不同的属性值被单独排序，然后，利用固定长度的窗口顺序扫描每一个记录序列，并在窗口内部对记录进行匹配操作。最后将多个属性上的匹配结果合并得到最后的实体解析结果中。假设窗口大小为l，记录数目为n，该方法能够将实体解析的代价从$O(n^2)$降至$O(l \quad n)$，这样一来，在实际应用中将会大大提升实体解析的效率。然而，在保证实体解析精度的情况下，l的最坏情况是n，因此算法的最坏时间代价仍然是$O(n^2)$。

Andrew McCallum等[38]利用数据分块处理的思想，借助一个代价不高的距离度量来有效地将数据分成重叠的子集。该方法首先将记录按照某些属性值的不同分为独立的块，然后在每个块内单独运行聚类算法，最后把每个块上的聚类结果合并得到实体解析结果中。该方法降低了每次调用聚类算法的时间代价，整体上提升了基于聚类方法的实体解析算法的效率。

甄灵敏等[39]针对关系数据库中实体解析效率问题，提出在基于分块技术的基础上采用信息增益方法和概率统计方法来计算记录属性的权重，该权重代表当前属性在记录中的重要性。通过将各个属性的权重分别计算以充分反映关键属性的重要性，是一种更符合现实的情况，因此这不仅提升了解析效率，而且解析的准确性也没有受到影响。

（2）自动学习

Hung-sik Kim等[40]针对数据规模比较大的情况，提出一个迭代的局部敏感Hash算法（Locality-Sensitive Hashing，LSH），以实现快速、精确的分块目的。由于该算法能动态合并基于LSH的Hash表，因此能对数据进行快速分块。重要的是，作者还给出了在解析速度上具有一定优越性的对应解析算法，因而能较好地提升解析效率。

Rares Vernica等[41]研究如何有效并行地执行实体解析，提出了利用云计算环境（Map Reduce）来加速大规模数据上的实体解析效率。由于在云计算环境基础上提出了一个基于数据分块计算思想的3阶段方法，因此可以以每个阶段为基础来探索若干解决方案，为高效的实体解析过程提供了新的思路。Mikhail Bilenko等[42]研究引入一个自适应框架来自动地学习能保证效率和准确性的分块函数。由于提出了两种基于谓词的可学习分块函数方法，并提供一个学习算法来训练他们，因此这种基于机器学习的自适应数据分块策略能较好地提升解析效率。

（3）分块方法比较

Rohan Baxter等[43]比较全面地综述了实体解析方法中的各种数据分块策略。将二元模型索引（Bigram Indexing）和Canopy聚类方法与标准的传统分块算法和近邻排序算法（Sorted-neighbourhood Blocking）方法进行比较。结果表明，由

于二元模型索引和Canopy聚类方法能提供可扩展的分块方法，因此有潜在提升速度和提高精度的可能性。

Toralf Kirsten等[44]对两种实际中经常用到的数据分块方法进行了形式化描述和对比分析。其中，一种是利用简单的策略（例如随即选取的Hash函数）将数据划分块，另一种是利用某些语义信息（例如基于属性值的描述性规则）将数据划分块。在对比中，从实体解析的时间效率方面来看，后一种方法具有明显的优势。然而，在实际应用中要找到具有适合语义信息的规则是非常困难的，有时甚至是不存在的。

4.2基于字符串相似

大多数应用在比较过程中都假设属性值是字符串，因此如何探索两个字符串在字符级别上、子字符串级别上的差异，并设计出有效的字符串相似算法，是需要考虑的重要方面。

（1）字符串特征

Nick Koudas等[45]较早地提出了针对字符串相似的实体解析优化问题。由于通过在大数据库上部署弹性的多个属性的字符串匹配方案，因此能给出初步的优化算法，但该算法使用了不能被文本方式捕捉的语义等价信息。Chaudhuri, S.等[46]对关系数据上基于字符串相似匹配的实体解析问题作进一步的抽象，提出了“相似连接”和“相似查询”操作，并将其作为数据库的一个基本操作来研究。

Chuan Xiao等[47]针对相似连接问题，提出将字符串的相似计算问题转化为集合的相似连接问题，并提出一个集合的相似连接操作算法。由于结合了基于字符串前缀、后缀的过滤方法，因此提出的方法能利用顺序信息避免对所有可能的记录对进行相似性计算，从而提升了基于相似连接的解析方法的效率。Panagiotis Papapetrou等[48]针对变长字符串，使用预先计算的对齐分（Alignment Scores），提出了基于变长字符串搜索的方法来解决长字符的相似查询问题，提升了属性值为长字符串情形的实体解析效率。

（2）n-gram（n元字符串）

Chen Li等[49-51]研究了基于n-gram的近似字符串匹配问题，其基本思想是在字符串上建立n-gram索引，将字符串之间的距离转化为对应n-gram交集的数量，然后基于n-gram的集合语义给出高效的相似连接算法，提升了实体解析效率。Behm, A.等 [52] 针对索引占用空间大的问题，提出了利用倒排索引来加速相似查询的方法。由于该方法基于丢弃字符串列表和组合相关列表来缩减索引空间，进而能维持有效的查询处理，因此提升了实体解析效率。

邱越峰等[53]提出了一种高效的基于n-gram的聚类算法，在聚类过程中，采用优先队列算法来准确地聚类相似重复记录，并以大量翔实的实验数据证明了此种解析方法的合理性和高效性。由于该算法能适应常见的拼写错误，如插入、删除、替换、交换和单词交换，因而有较好的解析效率，而且复杂度仅为$O(n)$。

4.3基于其他方法

在实体解析过程中如能有效地考虑其他一些重要信息，将能大大降低数据处理的时间和空间复杂度，进而提升解析效率。这方面的信息包括：图形处理器实体随时间演化的特性、大数据环境中的数据噪声、人机混合方法和大数据工具方法等。

Michael D. Lieberman 等[54]研究了高维数据上的实体解析问题，提出了一种基于图形处理器的相似联合算法，LSS 算法。由于利用了哈希技术，并结合图形处理器特性给出了两种基本的排序和检索数据操作对应的高效实现方法，因此该算法非常适合高维数据上的相似联合解析操作。

燕彩蓉等[55]基于 Map Reduce 编程模型，提出了一种迭代的并行处理框架。它采用面向学习的分类方法对实体进行解析，根据属性相似的传递性，并结合函数式语言的本身特性，对记录进行高效聚合。由于 Map Reduce 编程模型非常适合于实体解析过程一体化处理，因此作者提出的并行处理框架具有编程快捷、运行高效等特点，而且数据分区和并行处理技术避免了大量连接引发的内存溢出问题。燕彩蓉等[56]提出了一种机器计算与众包（Crowdsourcing）相结合的实体解析方法。该方法首先采用 MapReduce 并行计算框架排除不可能匹配的记录对，进而减少人类智能任务的数量，然后由人工进行确定性标注。此外，为了支持隐私保护，在众包计算时提出了基于角色的访问控制模型和重要信息隐藏策略。由于作者采用的人机结合方法充分利用了机器和人工处理的优势，因此解析过程中的高效率和高精度能较好地得到保障，并且能有效避免信息泄漏问题。

王宁等[57]针对大数据环境下传统的实体解析算法在效率、质量，特别是在抗噪声能力方面的表现并不理想的问题，提出了一种两层相关性聚类算法（Two-Tiered）。由于该算法基于相关性聚类（Correlation Clustering），且引入能有效定义节点和类之间关联程度的结点的邻居关系，因而提出的算法在计算代价、抗噪声能力和可扩展性方面均优于传统算法。

杨丹等[58]研究如何对数据空间中具有时间信息的实体进行解析，提出了一个四阶段的以时间为中心的集合实体解析策略（Time-Centered Collective Entity Resolution，T-CER），它基于以时间为基础的聚类算法（Time-based Clustering，T-Clustering）。在实体解析过程的不同阶段，T-CER 都考虑了时间信息所起的作用，并使用时间约束对解析结果进行检查。由于将数据的异构性和随时间演化的特性结合一起来考虑，因此提出的解析方法更具可行性和有效性。

4.4研究方法分析比较

以效率为目标的实体解析过程，主要关注相似重复记录间比较过程的效率。主要研究方法的比较情况，如表 2 所示：

表 2 以效率为目标的主要研究方法的比较

以效率为目标的实体解析（主要关注相似重复记录间比较过程的效率）		
采用的方法	**优点**	**缺点**
基于分块：①属性值；②自动学习和③分块方法比较	①不但有效地压缩了特征属性的维数，而且获得了组内的记录代表，为后面的高效准确解析打下了基础；②极大地减少比较计算的次数，从而在一定程度上降低了计算复杂度；③解析过程所需内存少，从而能够有效地实现对大量相似重复记录进行解析的目的	①效率在很大程度上取决于所选的键值；②选择键值通常依赖于数据所属领域的领域知识，因此需要对该领域具有深刻了解的专家的参与，这导致了方法自动化程度的降低和结果不确定性的增加；③如果键值选取不合适则会导致大量的重复数据被分到不同的子集合当中，这导致匹配数量的下降；④可能影响解析结果的完整性
基于字符串相	①使用的字符串算法比较成熟可	①不同的字段相似性计算方法往

似： ①字符串特征和②n 元字符串等	行；②可以很好地处理字符拼写错误情形；③具有很好的可伸缩性；④能有效解决属性和记录之间相似性是一种复杂非线性关系的问题	往对特定的字符串类型特别有效；②由于属性的相似性和记录的相似性之间是一种非线性的映射关系，因此把所有的属性值合并成一条长字符串或者简单地利用属性相似性的加权和来计算记录相似性的方法是不可取的
基于其他方法： ①图形处理器；②实体随时间演化的特性；③大数据环境中数据噪声；④人机混合方式和⑤大数据工具等	①充分利用了相应的特性来设计较优的匹配函数；②计算速度较快	①这些方法各有所长，但没有一种适用于所有数据集的方法，即不具有通用性；②减少了人为因素的影响；③可伸缩性差、自适应差

5 结论与展望

针对关系数据库中的实体解析技术，现有的工作主要在精度和效率两方面展开研究，力求在精度和效率之间找到一种合适、折衷的解析策略。尽管现有的研究工作设法从整体上改进实体解析技术，但适应于大数据环境的实体解析技术比较缺乏，尤其是在数据源的动态演化、异构性和非精确字符串匹配等方面。其中，这包括随时间变化的动态数据的实体解析，大规模的身份管理、隐私和查询驱动的实体解析以及主动学习和以众包为基础的实体解析。此外，基于图形来进行推理并解析的需求尽管超出了当前研究的理论应用，但意味着它也是一个可行的解决方案。特别地，基于增量式和基于分布式的两种解析策略能显著提高解析精度和提升解析效率，同时具有较好的可扩展性和高效性。

伴随着应用规模的不断扩大、数据量的急剧增长、数据关系的日益复杂化以及数据处理要求的不断提高。传统上实施一对一的记录比较过程往往不是最佳的方案，因为这需要大量的解析时间，从而难以满足效率要求，更难以胜任复杂的大数据环境。鉴于此，笔者认为未来实体解析技术还存在 3 个开放的研究方向

（1）面向数据记录的动态演化。一些应用中涉及的复杂数据记录会频繁更新，例如互联网信息和社会网络上的信息。因此，如何在更新频繁的动态复杂数据记录集上进行快速、有效的实体解析，是实体解析技术需要面对的主要挑战。

（2）面向数据记录的集成。对异构、海量的数据源进行数据抽取、清洗与整合是有效利用这些数据的前提。伴随而来的是数据记录间的不确定数据、结构不一致和模式匹配问题。因此，如何在这些情况下准确解析出描述同一实体的多个数据记录是实体解析技术需要面对的主要挑战。

（3）面向非精确字符串匹配。数据记录间的比较是一种计算复杂度很高的过程，而且由于匹配的记录对数量往往远少于不匹配的记录对数量，因而绝大部分比较过程浪费在不匹配的记录对之间。因此，如何研究出一些基础性方法，例如，非精确字符串匹配方法以及字符匹配过程中的最优过滤器选择等，以便能在匹配的准确性和完整性得到保障的同时尽量减少需要比较的记录数目，是实体解析技术需要面对的主要挑战。

作者贡献声明：

高广尚：研究过程实施，进行具体文献调研、分析与论文撰写、论文最终版本修订。
张智雄：提出研究思路。

（通信作者：高广尚 E-mail: gaoguangshang@mail.las.ac.cn)

6 参考文献

[1] NEWCOMBE H B，KENNEDY J M，AXFORD S，et al. Automatic Linkage of Vital Records [J]. Science, 1959, 130（3381）:954-959.

[2] FELLEGI I P，SUNTER A B. A theory for record linkage [J]. Journal of the American Statistical Association, 1969, 64（328）:1183-1210.

[3] NEWCOMBE H B，KENNEDY J M. Record linkage: making maximum use of the discriminating power of identifying information [J]. Commun ACM, 1962, 5（11）:563-566.

[4] HERN M A，#225，NDEZ，et al. The merge/purge problem for large databases [C] // in: Proceedings of the 1995 ACM SIGMOD international conference on Management of data, San Jose, California, USA. 223807: ACM, 1995: 127-138.

[5] SARAWAGI S，BHAMIDIPATY A. Interactive deduplication using active learning [C] // in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada. 775087: ACM, 2002: 269-278.

[6] DONG X，HALEVY A，MADHAVAN J. Reference reconciliation in complex information spaces [C] // in: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Baltimore, Maryland. 1066168: ACM, 2005: 85-96.

[7] TEJADA S，KNOBLOCK C A，MINTON S. Learning object identification rules for information integration [J]. Inf Syst, 2001, 26（8）:607-633.

[8] PETER C. Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection [M]. springer. 2012.

[9] ELMAGARMID A K，IPEIROTIS P G，VERYKIOS V S. Duplicate Record Detection: A Survey [J]. IEEE Trans on Knowl and Data Eng, 2007, 19（1）:1-16.

[10]WINKLER W E. Overview of record linkage and current research directions [R]. Citeseer: BUREAU OF THE CENSUS, 2006.

[11]BENJELLOUN O，GARCIA-MOLINA H，MENESTRINA D，et al. Swoosh: a generic approach to entity resolution [J]. The VLDB Journal, 2009, 18（1）:255-276.

[12]BHATTACHARYA I，GETOOR L. Collective entity resolution in relational data [J]. ACM Trans Knowl Discov Data, 2007, 1（1）:5.

[13]MANNING C D，RAGHAVAN P，SCH\ H，et al. Introduction to Information Retrieval [M]. Cambridge University Press. 2008: 496.

[14]ARASU A，G M，#246，et al. On active learning of record matching packages [C] // in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, USA. 1807252: ACM, 2010: 783-794.

[15]刘骏豪，孙晶莹. 2011 年德国人口普查中的新技术——记录连接 [J]. 中国统计, 2011,（11）:38-39.

[16]谭明超，刁兴春，曹建军. 实体分辨研究综述 [J]. 计算机科学, 2014, 41（4）:9-12, 20. （TAN Ming-chao, DIAO Xing-chun,CAO Jian-jun, Survey on Entity Resolution .Computer Science. 2014, 41（4）:9-12, 20.）

[17]M LLER H，FREYTAG J-C. Problems, methods, and challenges in comprehensive data cleansing [M]. Professoren des Inst. Für Informatik. 2005.

[18]Record Linkage in Large Data Sets[EB/OL].[2014-12-02]. http://www.dani-sola.com/record-linkage-in-large-data-sets/.

[19]REITER J. Data Quality and Record Linkage Techniques [J]. Journal of the American Statistical Association, 2008, 103（482）:881-881.

[20]WINKLER W E. Methods for record linkage and bayesian networks [R]. Statistical Research Division, US Census Bureau, Washington, DC, 2002.
[21]WHANG S E，GARCIA-MOLINA H. Entity resolution with evolving rules [C] // in: Proceedings of the VLDB Endowment,, Singapore, 2010: 1326-1337.
[22]WHANG S E，GARCIA-MOLINA H. Incremental entity resolution on rules and data [J]. The VLDB Journal, 2014, 23（1）:77-102.
[23]WHANG S E，GARCIA-MOLINA H. Developments in generic entity resolution [J]. IEEE Data Engineering Bulletin, 2011, 13（11）:24-30.
[24]WHANG S E，MENESTRINA D，KOUTRIKA G，et al. Entity resolution with iterative blocking [C] // in: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, Providence, Rhode Island, USA. 1559870: ACM, 2009: 219-232.
[25]GRUENHEID A，DONG X L，SRIVASTAVA D. Incremental Record Linkage [C] // in: Proceedings of the VLDB Endowment, Hangzhou, ChinaVLDB Endowment, 2014: 20-12.
[26]SARAWAGI S，DESHPANDE V S，KASLIWAL S. Efficient top-k count queries over imprecise duplicates [C] // in: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, Saint Petersburg, Russia. 1516413: ACM, 2009: 450-461.
[27]DEZ M A H，STOLFO S J. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem [J]. Data Min Knowl Discov, 1998, 2（1）:9-37.
[28]MATHIEU C，SANKUR O，SCHUDY W. Online correlation clustering [J]. arXiv preprint arXiv:10010920, 2010, 12（3）:21-36.
[29]CHARIKAR M，CHEKURI C，FEDER T，et al. Incremental clustering and dynamic information retrieval [C] // in: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing, ACM, 1997: 626-635.
[30]AGGARWAL C C，HAN J，WANG J，et al. A framework for clustering evolving data streams [C] // in: Proceedings of the 29th international conference on Very large data bases - Volume 29, Berlin, Germany. 1315460: VLDB Endowment, 2003: 81-92.
[31]SINGLA P，DOMINGOS P. Collective object identification [C] // in: Proceedings of the 19th international joint conference on Artificial intelligence, Edinburgh, Scotland. 1642589: Morgan Kaufmann Publishers Inc., 2005: 1636-1637.
[32]CHRISTEN P. Automatic record linkage using seeded nearest neighbour and support vector machine classification [C] // in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA. 1401913: ACM, 2008: 151-159.
[33]楼俊杰，徐从富，郝春亮. 基于马尔科夫逻辑网络的实体解析改进算法 [J]. 计算机科学, 2010, （08）:243-247. （LOU Jun-jie XU Cong-fu HAO Chun-liang. Improvement of Entity Resolution Based on Markov Logic Networks. Computer Science,2010, （08）:243-247.）
[34]CHAUDHURI S，GANTI V，XIN D. Mining document collections to facilitate accurate approximate entity matching [J]. Proc VLDB Endow, 2009, 2（1）:395-406.
[35]LIANGCAI S，BO L，WEIYI M. A Latent Topic Model for Complete Entity Resolution [C] // in: Data Engineering, 2009 ICDE '09 IEEE 25th International Conference on, 2009: 880-891.
[36]RASTOGI V，DALVI N，GAROFALAKIS M. Large-scale collective entity matching [C] // in: Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, Washington. USA: VLDB, 2011: 208-218.
[37]GETOOR L，MACHANAVAJJHALA A. Entity resolution: theory, practice & open challenges [J]. Proc VLDB Endow, 2012, 5（12）:2018-2019.
[38]MCCALLUM A，NIGAM K，UNGAR L H. Efficient clustering of high-dimensional data sets with application to reference matching [C] // in: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts,USA. USA: ACM, 2000: 169-178.
[39]甄灵敏，杨晓春，王斌，et al. 基于属性权重的实体解析技术 [J]. 计算机研究与发展,

2013, （S1）:281-289. （Zhen Lingmin,Yang Xiaochun,Wang Bin,and Ahmed A Hussein. An Entity Resolution Approach Based on Attributes Weights. Journal of Computer Research and Development, 2013, （S1）:281-289.）
[40]KIM H-S，LEE D. HARRA: fast iterative hashed record linkage for large-scale data collections [C] // in: Proceedings of the 13th International Conference on Extending Database Technology, Lausanne, Switzerland. USA: ACM, 2010: 525-536.
[41]VERNICA R，CAREY M J，LI C. Efficient parallel set-similarity joins using MapReduce [C] // in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, USA. 1807222: ACM, 2010: 495-506.
[42]BILENKO M，KAMATH B，MOONEY R J. Adaptive Blocking: Learning to Scale Up Record Linkage [C] // in: Data Mining, 2006 ICDM '06 Sixth International Conference on, USA. USA: IEEE, 2006: 87-96.
[43]BAXTER R，CHRISTEN P，CHURCHES T. A Comparison of Fast Blocking Methods for Record Linkage [C] // in: the First Workshop on Data Cleaning, Record Linkage and Object Consolidation, KDD, Washington, DC. USA: KDD, 2003: 25-27 %&.
[44]KIRSTEN T，KOLB L，HARTUNG M，et al. Data partitioning for parallel entity matching [J]. arXiv preprint arXiv:10065309, 2010, 10（4）:20-29.
[45]KOUDAS N，MARATHE A，SRIVASTAVA D. Flexible string matching against large databases in practice [C] // in: Proceedings of the Thirtieth international conference on Very large data bases - Volume 30, Toronto, Canada. 1316782: VLDB Endowment, 2004: 1078-1086.
[46]CHAUDHURI S，GANTI V，KAUSHIK R. A Primitive Operator for Similarity Joins in Data Cleaning [C] // in: Data Engineering, 2006 ICDE '06 Proceedings of the 22nd International Conference on, 2006: 5-5.
[47]XIAO C，WANG W，LIN X，et al. Efficient similarity joins for near duplicate detection [C] // in: Proceedings of the 17th international conference on World Wide Web, Beijing, China. 1367516: ACM, 2008: 131-140.
[48]PAPAPETROU P，ATHITSOS V，KOLLIOS G，et al. Reference-based alignment in large sequence databases [J]. Proc VLDB Endow, 2009, 2（1）:205-216.
[49]LI C，LU J，LU Y. Efficient Merging and Filtering Algorithms for Approximate String Searches [C] // in: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. 1547171: IEEE Computer Society, 2008: 257-266.
[50]LI C，WANG B，YANG X. VGRAM: improving performance of approximate queries on string collections using variable-length grams [C] // in: Proceedings of the 33rd international conference on Very large data bases, Vienna, Austria. USA: VLDB Endowment, 2007: 303-314.
[51]YANG X，WANG B，LI C. Cost-based variable-length-gram selection for string collections to support approximate queries efficiently [C] // in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vancouver, Canada. 1376655: ACM, 2008: 353-364.
[52]BEHM A，SHENGYUE J，CHEN L，et al. Space-Constrained Gram-Based Indexing for Efficient Approximate String Search [C] // in: Data Engineering, 2009 ICDE '09 IEEE 25th International Conference on Data Engineering, USA. USA: IEEE, 2009: 604-615.
[53]邱越峰，田增平．一种高效的检测相似重复记录的方法 [J]. 计算机学报, 2001, 24（1）:69-77. （QIU Yue-Feng TIAN Zeng-Ping JI Wen-Yun ZHOU Ao-Ying. An Efficient Approach for Detecting Approximately Duplicate Database Records. Chinese Journal of Computers. , 2001, 24（1）:69-77.）
[54]LIEBERMAN M D，SANKARANARAYANAN J，SAMET H. A Fast Similarity Join Algorithm Using Graphics Processing Units [C] // in: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, USA. USA: IEEE Computer Society, 2008: 1111-1120.
[55]燕彩蓉，万永权．并行实体解析与记录聚合模型 [J]. 小型微型计算机系统, 2013,

（ 08 ） :1843-1847. （ YAN Cai-rong,Wan Yong-quan. Journal of Chinese Systems. Parallel Entity Resolution and Record Aggregation Model.）
[56]燕彩蓉，张洋舜，徐光伟. 支持隐私保护的众包实体解析 [J]. 计算机科学与探索, 2014, （07）:802-811. （YAN Cairong, ZHANG Yangshun, XU Guangwei. Crowdsourcing entity resolution with privacy protection. Journal of Frontiers of Computer Science and Technology, 2014, 8(7)：802-811.）
[57]王宁，李杰. 大数据环境下用于实体解析的两层相关性聚类方法 [J]. 计算机研究与发展， 2014, （ 09 ）:2108-2116. （ Wang Ning and Li Jie. Two-Tiered Correlation Clustering Method for Entity Resolution in Big Data. Journal of Computer Research and Development, 2014, （09）:2108-2116.）
[58]杨丹，申德荣，于戈，et al. 数据空间中时间为中心的集合实体识别策略 [J]. 计算机科学与探索, 2012, 6（11）:974-984. （YANG Dan, SHEN Derong, YU Ge, et al. Time-centered collective entity resolution strategy in dataspace. Journal of Frontiers of Computer Science and Technology, 2012, 6(11)：974-984.）